Expolortary Data Analysis

(EDA)

Asad Raza Virk

2024-09-18

Table of contents

1 Elements of Structured Data

- Data comes from many sources: sensor measurements, events, text, images, and videos.
- The Internet of Things (IoT) is spewing out streams of information. Much of this data is unstructured: images are a collection of pixels, with each pixel containing RGB (red, green, blue) color information.
- Texts are sequences of words and non-word characters, often organized by sections, subsections, and so on.
- Clickstreams are sequences of actions by a user interacting with an app or a web page.
- A major challenge of data science is to harness this torrent of raw data into actionable information.
- To apply the statistical concepts , unstructured raw data must be processed and manipulated into a structured form.

One of the commonest forms of structured data is a table with rows and columns—as data might emerge from a relational database or be collected for a study

1.1 Two basic types of Structured Data

1. Numeric
    - continuous such as wind speed or time duration
    - discrete
      such as the count of the occurrence of an event

2. Categorical (takes only fixed set of values)
    - Binary Binary data is an important special case of categorical data that takes on only one of two values, such as 0/1, yes/no, or true/false
    - Ordinal Ordinal data in which the categories are ordered; an example of this is a numerical rating (1, 2, 3, 4, or 5)

For the purposes of data analysis and predictive modeling, the data type is important to help determine the type of visual display, data analysis, or statistical model.

Data science software, such as R and Python, uses these data types to improve computational performance. More important, the data type for a variable determines how software will handle computations for that variable.

1.2 Key Terms for Data Types

1. Numeric Data that are expressed on a numeric scale.
    - Continuous Data that can take on any value in an interval. (Synonyms: Interval, float, numeric)
    - Discrete Data that can take on only integer values, such as counts. (Synonyms: integer, count)
2. Categorical Data that can take on only a specific set of values representing a set of possible categories. (Synonyms: enums, enumerated, factors, nominal)
    - Binary A special case of categorical data with just two categories of values, e.g., 0/1, true/false. (Synonyms: dichotomous, logical, indicator, boolean)
    - Ordinal Categorical data that has an explicit ordering. (Synonym: ordered factor)

- Key Ideas

- Data is typically classified in software by type.
- Data types include numeric (continuous, discrete) and categorical (binary, ordinal).
- Data typing in software acts as a signal to the software on how to process the data.

2 Rectangular and Non-Rectangualr Data

2.1 Rectangular Data

- The typical frame of reference for an analysis in data science is a rectangular data object, like a spreadsheet or database table.
- Rectangular data is the general term for a two-dimensional matrix with rows indicating records (cases) and columns indicating features (variables)
- Data frame is the specific format in R and Python.
- The data doesn't always start in this form: unstructured data (e.g., text) must be processed and manipulated so that it can be represented as a set of features in the rectangular data
- Data in relational databases must be extracted and put into a single table for most data analysis and modeling tasks.

2.1.1 Key Terms for Rectangular Data

1. Data frame Rectangular data (like a spreadsheet) is the basic data structure for statistical and machine learning models.
2. Feature A column within a table is commonly referred to as a feature. Synonyms: attribute, input, predictor, variable
3. Outcome Many data science projects involve predicting an outcome—often a yes/no outcome. The features are sometimes used to predict the outcome in an experiment or a study. Synonyms: dependent variable, response, target, output
4. Records A row within a table is commonly referred to as a record. Synonyms: case, example, instance, observation, pattern, sample

2.1.2 Data Frames and Indexes

Traditional database tables have one or more columns designated as an index, essentially a row number. This can vastly improve the efficiency of certain database queries.

In Python, with the pandas library, the basic rectangular data structure is a DataFrame object. By default, an automatic integer index is created for a DataFrame based on the order of the rows.

In pandas, it is also possible to set multilevel/hierarchical indexes to improve the efficiency of certain operations.

> Terminology Diferences Terminology for rectangular data can be confusing. Statisticians and data scientists use different terms for the same thing. For a statistician, predictor variables are used in a model to predict a response or dependent variable. For a data scientist, features are used to predict a target. One synonym is particularly confusing: computer scientists will use the term sample for a single row; a sample to a statistician means a collection of rows.

2.2 Non-Rectangular Data Structures

There are other data structures besides rectangular data.

- Time series data records successive measurements of the same variable. It is the raw material for statistical forecasting methods, and it is also a key component of the data produced by devices—the Internet of Things.
- Spatial data structures, which are used in mapping and location analytics, are more complex and varied than rectangular data structures. In the object representation, the focus of the data is an object (e.g., a house) and its spatial coordinates. The field view, by contrast, focuses on small units of space and the value of a relevant metric (pixel brightness, for example).
- Graph (or network) data structures are used to represent physical, social, and abstract relationships. For example, a graph of a social network, such as Facebook or LinkedIn, may represent connections between people on the network. Distribution hubs connected by roads are an example of a physical network. Graph structures are useful for certain types of problems, such as network optimization and recommender systems.

  > Graphs in Statistics In computer science and information technology, the term graph typically refers to a depiction of the connections among entities, and to the underlying data structure. In statistics, graph is used to refer to a variety of plots

and visualizations, not just of connections among entities, and the term applies only to the visualization, not to the data structure.

3 Estimates of Location of Data

- Variables with measured or count data might have thousands of distinct values.
- A basic step in exploring your data is getting a "typical value" for each feature (variable): an estimate of where most of the data is located (i.e., its central tendency).

3.1 Key Terms for Estimates of Location

- Mean
  - The sum of all values divided by the number of values.
    - Synonym
      - average
- Weighted mean
  - The sum of all values times a weight divided by the sum of the weights.
    - Synonym
      - weighted average
- Median
  - The value such that one-half of the data lies above and below.
    - Synonym
      - 50th percentile
- Percentile
  - The value such that P percent of the data lies below.
    - Synonym
      - quantile
- Weighted median
  - The value such that one-half of the sum of the weight
- Trimmed mean
  - The average of all values after dropping a fixed number of extreme values.

        * Synonym
            · truncated mean

- Robust
    - Not sensitive to extreme values.
        * Synonym
            · resistant

- Outlier
    - A data value that is very different from most of the data.
        * Synonym
            · extreme value

Metrics and Estimates

Statisticians often use the term >estimate for a value calculated from the data at hand, to draw a distinction between what we see from the data and the theoretical true or exact state of affairs. Data scientists and >business analysts are more likely to refer to such a value as a metric

```
# Import the important Statistical libraries
import numpy as np # Fundamental package for numerical computing
import pandas as pd # For Data manipulation and analysis
from scipy import stats # For Advanced scientific computing and statistical analysis
import statsmodels.api as sm # For Statistical modeling and hypothesis testing
import seaborn as sns # For Data visualization for understanding statistical distributions
import matplotlib.pyplot as plt # For Data visualization for understanding statistical distributions
import plotly as plt # For Data visualization for understanding statistical distributions
```

```
# Load the US Health Insurance Dataset downlaoded from Kaggle
df = pd.read_csv ('./data_sets/insurance.csv')
```

3.2 Mean

3.2.1 Mean

- The mean is the sum of all values divided by the number of values.

```
# Mean with pandas library
print ("The Mean of Charges is ", df['charges'].mean())
```

The Mean of Charges is 13270.422265141257

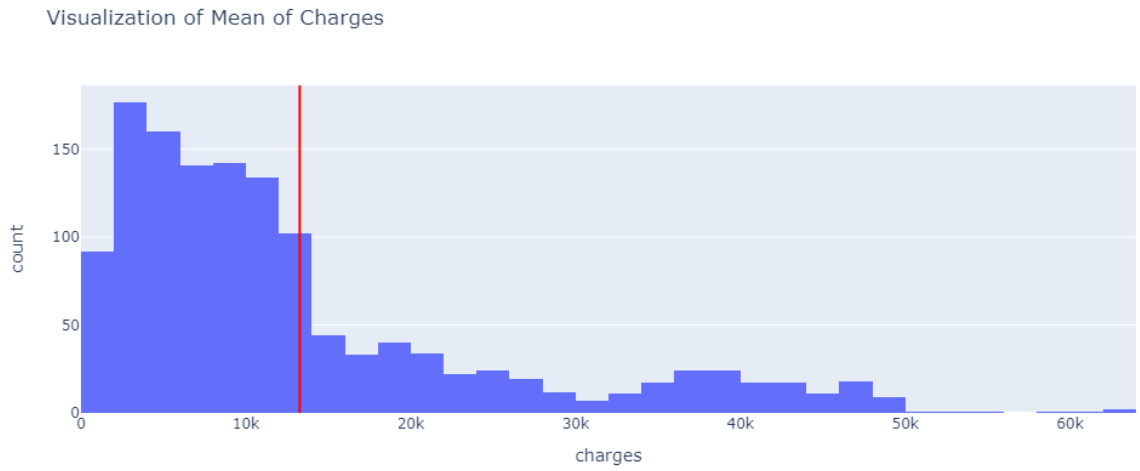


Figure 1: Mean

3.2.2 Trimmed Mean

- A variation of the mean which calculate the mean by dropping a fixed number of sorted values at each end and then taking an average of the remaining values
- A trimmed mean eliminates the influence of extreme values
- The trimmed mean is a robust measure of central tendency, as it is less affected by outliers than the mean
- The trimmed mean is calculated by first sorting the data in ascending order, then dropping a fixed number

- For example, in international diving the top score and bottom score from five judges are dropped, and the final score is the average of the scores from the three remaining judges. This makes it difficult for a single judge to manipulate the score, perhaps to favor their country's contestant.
- Trimmed means are widely used, and in many cases are preferable to using the ordinary mean

```
# import the library for trim mean
from scipy.stats import trim_mean
# Calculate the trimmed mean, trimming 10% from each end
trimmed_mean_charges = trim_mean(df['charges'], proportiontocut=0.1)
print("Trimmed Mean charges:", trimmed_mean_charges)
```

Trimmed Mean charges: 11076.019520008396

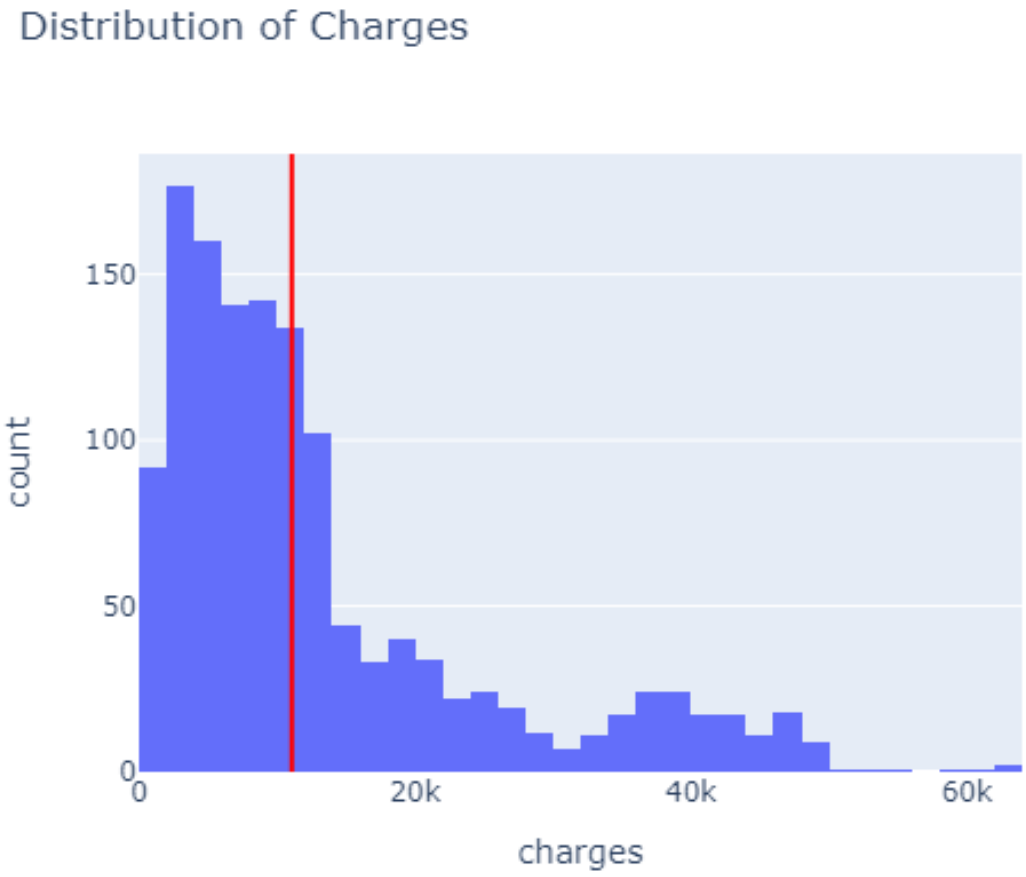
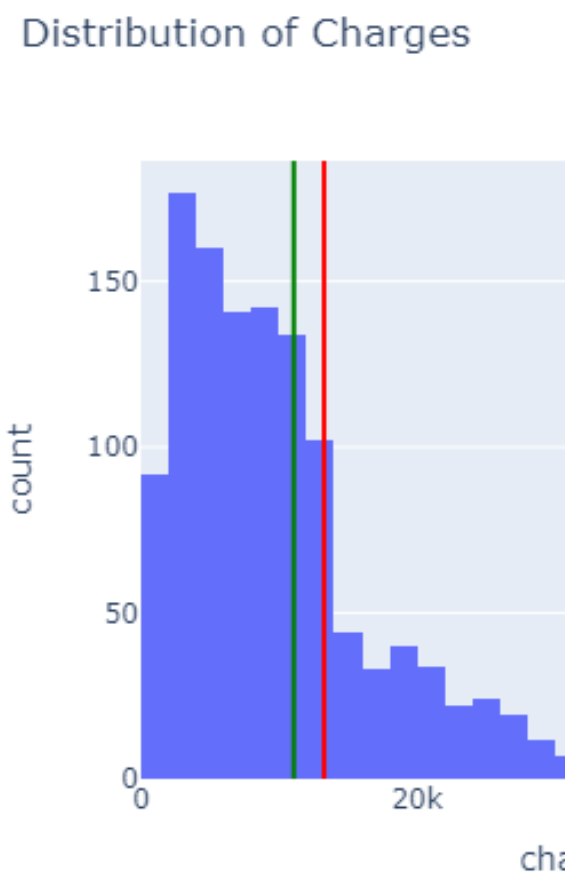


Figure 2: Trimmed Mean



Mean and Trimmed Mean (Mean = Red, Trimmed Mean = Greed)

3.3 Outliers

- outliers (extreme cases) could skew the results
- An outlier is any value that is very distant from the other values in a data set
- The exact definition of an outlier is somewhat subjective
- Outliers can be either high or low values
- Being an outlier in itself does not make a data value invalid or erroneous
- outliers are often the result of data errors such as mixing data of different units (kilometers versus meters) or bad readings from a sensor.
- When outliers are the result of bad data, the mean will result in a poor estimate of location, while the median will still be valid
- In any case, outliers should be identified and are usually worthy of further investigation.

Anomaly Detection

In contrast to typical data analysis, where outliers are sometimes informative and sometimes a nuisance, in anomaly detection the points of interest are the outliers, and the greater mass of data serves primarily to define the "normal" against which anomalies are measured.

The median is not the only robust estimate of location. In fact, a trimmed mean is widely used to avoid the influence of outliers. For example, trimming the bottom and top 10% (a common choice) of the data will provide protection against outliers in all but the smallest data sets. The trimmed mean can be thought of as a compromise between the median and the mean: it is robust to extreme values in the data, but uses more data to calculate the estimate for location.

Weighted mean is available with NumPy. For weighted median, we can use the specialized package wquantiles

3.3.1 Key Ideas

- The basic metric for location is the mean, but it can be sensitive to extreme values (outlier).
- Other metrics (median, trimmed mean) are less sensitive to outliers and unusual distributions and hence are more robust

4 Estimates of Variability

4.1 Variability

- Location is just one dimension in summarizing a feature
- A second dimension, variability, also referred to as dispersion, measures whether the data values are tightly clustered or spread out.
- At the heart of statistics lies variability:
  - measuring it
  - reducing it
  - distinguishing random from real variability
  - identifying the various sources of real variability
  - making decisions in the presence of it
- Variability is often measured using the range, interquartile range (IQR), variance, or standard deviation

4.1.1 Key Terms for Variability Metrics

- Deviations
  - The difference between the observed values and the estimate of location.
    - Synonyms
      - errors, residuals
- Variance
  - The sum of squared deviations from the mean divided by n – 1 where n is the number of data values.
    - Synonym
      - mean-squared-error (MSE)

Step-by-Step Example of Sample Variance Calculation:

1. Data Points: {3, 5, 7, 9, 11}

2. Calculate the Mean:
   Mean ($\bar{x}$) = (3 + 5 + 7 + 9 + 11) / 5 = 7

3. Subtract the Mean from Each Data Point (Deviations):
   (3 - 7) = -4, (5 - 7) = -2, (7 - 7) = 0, (9 - 7) = 2, (11 - 7) = 4
   Deviations: {-4, -2, 0, 2, 4}

4. Square Each Deviation:
   (-4)^2 = 16, (-2)^2 = 4, (0)^2 = 0, (2)^2 = 4, (4)^2 = 16
   Squared Deviations: {16, 4, 0, 4, 16}

5. Sum of Squared Deviations:
   16 + 4 + 0 + 4 + 16 = 40

6. Divide by n-1 (n=5):
   Variance = 40 / (5 - 1) = 40 / 4 = 10

Sample Variance = 10

Figure 3: Example

- Standard deviation
  - The square root of the variance.
    - Synonyms
      - root mean squared error (RMSE)
- Mean absolute deviation
  - The mean of the absolute values of the deviations from the mean.
    - Synonyms
      - l1-norm, Manhattan norm

Step-by-Step Example of Mean Absolute Deviation (MAD) Calculation:

Data points: {2, 4, 6, 8, 10}

1. Calculate the Mean:
   Mean ($\bar{x}$) = (2 + 4 + 6 + 8 + 10) / 5 = 6

2. Find the Absolute Deviations from the Mean:
   |2 - 6| = 4
   |4 - 6| = 2
   |6 - 6| = 0
   |8 - 6| = 2
   |10 - 6| = 4
   Absolute deviations: {4, 2, 0, 2, 4}

3. Calculate the Mean of the Absolute Deviations:
   MAD = (4 + 2 + 0 + 2 + 4) / 5 = 12 / 5 = 2.4

Mean Absolute Deviation (MAD) = 2.4

Figure 4: Example

- Median absolute deviation from the median
  - The median of the absolute values of the deviations from the median

Step-by-Step Example of Median Absolute Deviation (MAD) Calculation:

Data points: {2, 4, 6, 8, 10}

1. Find the Median of the Dataset:
   Median = 6 (the middle value in the sorted dataset)

2. Find the Absolute Deviations from the Median:
   |2 - 6| = 4
   |4 - 6| = 2
   |6 - 6| = 0
   |8 - 6| = 2
   |10 - 6| = 4
   Absolute deviations: {4, 2, 0, 2, 4}

3. Find the Median of the Absolute Deviations:
   Sorted absolute deviations: {0, 2, 2, 4, 4}
   Median of absolute deviations = 2

Median Absolute Deviation (MAD) = 2

Figure 5: Example

- Range
  - The difference between the largest and the smallest value in a data set.
- Order statistics
  - Metrics based on the data values sorted from smallest to biggest.
    - Synonym
      - ranks
- Percentile
  - The value such that P percent of the values take on this value or less and (100–P) percent take on this value or more.
    - Synonym
      - quantile

- Interquartile range
  - The difference between the 75th percentile and the 25th percentile.
    - Synonym
      - IQR

4.2 Standard Deviation and Related Estimates

- The most widely used estimates of variation are based on the differences, or deviations, between the estimate of location and the observed data.
- For a set of data {1, 4, 4}, the mean is 3 and the median is 4. The deviations from the mean are the differences: 1 – 3 = –2, 4 – 3 = 1, 4 – 3 = 1.
- These deviations tell us how dispersed the data is around the central value
- One way to measure variability is to estimate a typical value for these deviations.Averaging the deviations themselves would not tell us much—the negative deviations offset the positive ones. In fact, the sum of the deviations from the mean is precisely zero.
- Instead, a simple approach is to take the average of the absolute values of the deviations from the mean. In the preceding example, the absolute value of the deviations is {2 1 1}, and their average is (2 + 1 + 1) / 3 = 1.33. This is known as the mean absolute deviation
- The best-known estimates of variability are the variance and the standard deviation, which are based on squared deviations. The variance is an average of the squared deviations, and the standard deviation is the square root of the variance
- The standard deviation is much easier to interpret than the variance since it is on the same scale as the original data. Still, with its more complicated and less intuitive formula, it might seem peculiar that the standard deviation is preferred in statistics over the mean absolute deviation. It owes its preeminence to statistical theory: mathematically, working with squared values is much more convenient than absolute values, especially for statistical models.

  here is always some discussion of why we have n – 1 in the denominator in the variance formula, instead of n, leading into the concept of degreesof freedom. This distinction is not important since n is generally large enough that it won't make much difference whether you divide by n or n – 1. But in case you are interested, here is the story. It is based on the premise that you want to make estimates about a population, based

on a sample. If you use the intuitive denominator of n in the variance formula, you will underestimate the true value of the variance and the standard deviation in the population. This is referred to as a biased estimate. However, if you divide by n – 1 instead of n, the variance becomes an unbiased estimate. To fully explain why using n leads to a biased estimate involves the notion of degrees of freedom, which takes into account the number of constraints in computing an estimate. In this case, there are n – 1 degrees of freedom since there is one constraint: the standard deviation depends on calculating the sample mean. For most problems, data scientists do not need to worry about degrees of freedom.

- Neither the variance, the standard deviation, nor the mean absolute deviation is robust to outliers and extreme values
- The variance and standard deviation are especially sensitive to outliers since they are based on the squared deviations.
- A robust estimate of variability is the median absolute deviation from the median or MAD
- Like the median, the MAD is not influenced by extreme val- ues.
- It is also possible to compute a trimmed standard deviation analogous to the trimmed mean > The variance, the standard deviation, the mean absolute deviation, and the median absolute deviation from the median are not equivalent estimates, even in the case where the data comes from a normal distribution. In fact, the standard deviation is always greater than the mean absolute deviation, which itself is greater than the median absolute deviation. Sometimes, the median absolute deviation is multiplied by a constant scaling factor to put the MAD on the same scale as the standard deviation in the case of a normal distribution. The commonly used factor of 1.4826 means that 50% of the normal distribution fall within the range ±MAD

The purpose of calculating deviations in the values is to understand how each individual value differs from the average (mean) value. This helps in several ways:

- Measure of Variability: Deviations provide a measure of the spread or variability in the data. Large deviations indicate that the charges are spread out over a wide range, while small deviations suggest that the charges are clustered closely around the mean.
- Identify Outliers: By examining the deviations, you can identify outliers or unusual data points that are significantly different from the mean.

- Statistical Analysis: Deviations are a fundamental component in various statistical analyses, such as calculating the variance and standard deviation, which are key measures of data dispersion.
- Data Visualization: Plotting the deviations can help visualize the distribution of the data, making it easier to identify patterns, trends, and anomalies.

In statistical analysis, variance and deviation are closely related concepts that measure the spread or dispersion of a dataset. Here's how they are related:

- Deviation:
  - Deviation refers to the difference between each data point and the mean of the dataset.
  - Deviations can be positive or negative, depending on whether the data point is above or below the mean.
  - The sum of deviations from the mean is always zero, as positive deviations cancel out negative deviations
- Variance:
  - Variance is a measure of how much the data points in a dataset vary from the mean.
  - It is calculated as the average of the squared deviations from the mean.
  - Squaring the deviations ensures that all values are positive and gives more weight to larger deviations.
  - The variance is always non-negative and is zero if all data points are identical
- Relationship:
  - Variance is essentially the mean of the squared deviations.
  - While deviations provide a measure of individual differences from the mean, variance provides a single value that summarizes the overall dispersion of the dataset.
  - Variance is used to calculate the standard deviation, which is the square root of the variance and provides a measure of dispersion in the same units as the original data.

In summary, deviations are the building blocks for calculating variance, and variance provides a comprehensive measure of the spread of the data based on these deviations.

This code will output a table with the columns for value, mean, deviation, and variance, and it will also plot the deviations. Adjust the df[‘charges’] list with your actual data if needed.

Value	Mean	Deviation	Variance	Squared Deviation
16884.924	13270.422265141257	3614.5017348587426	146542766.49354792	13064622.79129686
1725.5523	13270.422265141257	-11544.869965141257	146542766.49354792	133284022.51202069
4449.462	13270.422265141257	-8820.960265141257	146542766.49354792	77809339.99920091
21984.47061	13270.422265141257	8714.048344858744	146542766.49354792	75934638.55653541
3866.8552	13270.422265141257	-9403.567065141257	146542766.49354792	88427073.54860935
3756.6216	13270.422265141257	-9513.800665141256	146542766.49354792	90512403.0960422
8240.5896	13270.422265141257	-5029.832665141257	146542766.49354792	25299216.639322
7281.5056	13270.422265141257	-5988.916665141256	146542766.49354792	35867122.822006665
6406.4107	13270.422265141257	-6864.011565141256	146542766.49354792	47114654.76639292
28923.13692	13270.422265141257	15652.714654858744	146542766.49354792	245007476.0664297

Figure 6: Example

The variance is a single value that summarizes the overall dispersion of the dataset. It is not specific to individual data points but rather describes the dataset as a whole. Therefore, when you include the variance in the DataFrame, it is the same for every row because it represents the same overall measure of variability for the entire dataset.

If you want to include the variance in the DataFrame, it should be shown as a single value, not repeated for each data point. However, if you want to show the squared deviations (which contribute to the variance), you can include those instead.

In statistical analysis, the standard deviation and variance are both measures of the spread or dispersion of a dataset. They are closely related but differ in how they express this dispersion.

- Variance
- Definition: Variance measures the average squared deviations from the mean.

- Units: The units of variance are the square of the units of the original data. For example, if the data is in meters, the variance will be in square meters.
- Formula: Variance = Σ(xi -)^2 / (n - 1), where
  - xi is each data point,
  - is the mean of the dataset,
  - n is the number of data points, and
  - Σ denotes the sum of the squared deviations.
- Standard Deviation
- Definition: Standard deviation is the square root of the variance. It provides a measure of dispersion in the same units as the original data.
- Units: The units of standard deviation are the same as the units of the original data. For example, if the data is in meters, the standard deviation will also be in meters.
- Relationship Mathematical Relationship: The standard deviation is the square root of the variance.
- Interpretation:
  - Variance gives a measure of how data points spread out from the mean, but because it uses squared units, it can be less intuitive.
  - Standard deviation, being in the same units as the data, is often more interpretable and is commonly used to describe the spread of the data.
- Example If you have a dataset, the variance tells you the average of the squared differences from the mean, while the standard deviation tells you how much the data points typically deviate from the mean in the original units of the data.

4.3 Estimates Based on Percentile**

A different approach to estimating dispersion is based on looking at the spread of the sorted data.Statistics based on sorted (ranked) data are referred to as order statistics. - The most basic measure is the range: the difference between the largest and smallest numbers. The minimum and maximum values themselves are useful to know and are helpful in identifying outliers, but the range

is extremely sensitive to outliers and not very useful as a general measure of dispersion in the data. - To avoid the sensitivity to outliers, we can look at the range of the data after dropping values from each end Formally, these types of estimates are based on differences between percentiles - In a data set, the Pth percentile is a value such that at least P percent of the values take on this value or less and at least (100 – P) percent of the values take on this value or more. For example, to find the 80th percentile, sort the data. Then, starting with the smallest value, proceed 80 percent of the way to the largest value. Note that the median is the same thing as the 50th percentile. The percentile is essentially the same as a quantile, with quantiles indexed by fractions (so the .8 quantile is the same as the 80th percentile). - A common measurement of variability is the difference between the 25th percentile and the 75th percentile, called the interquartile range (or IQR). - Software can have slightly differing approaches that yield different answers - For very large data sets, calculating exact percentiles can be computationally very expensive since it requires sorting all the data values. Machine learning and statistical software use special algorithms, such as [Zhang-Wang-2007], to get an approximate percentile that can be calculated very quickly and is guaranteed to have a certain accuracy.

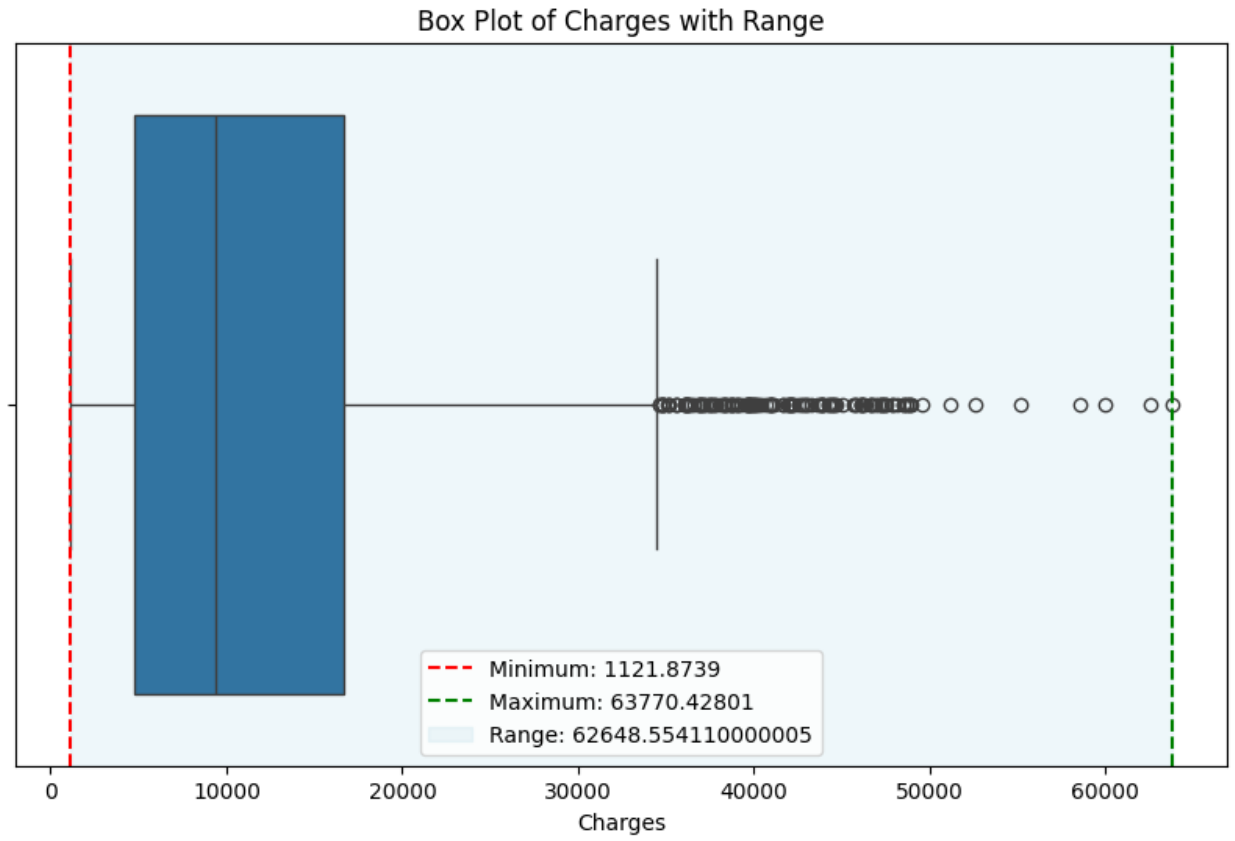


Figure 7: Range

Order statistics are metrics based on the data values sorted from smallest to largest. They provide insights into the distribution and spread of the data. Here are some common order statistics:

- Minimum: The smallest value in the dataset.
- Maximum: The largest value in the dataset.
- Median: The middle value when the data is sorted. If the dataset has an even number of observations, the median is the average of the two middle values.
- Quartiles: Values that divide the dataset into four equal parts.
  - First Quartile (Q1): The median of the lower half of the dataset (25th percentile).
  - Second Quartile (Q2): The median of the dataset (50th percentile).
  - Third Quartile (Q3): The median of the upper half of the dataset (75th percentile).
- Percentiles: Values that divide the dataset into 100 equal parts. For example, the 90th percentile is the value below which 90% of the data falls.

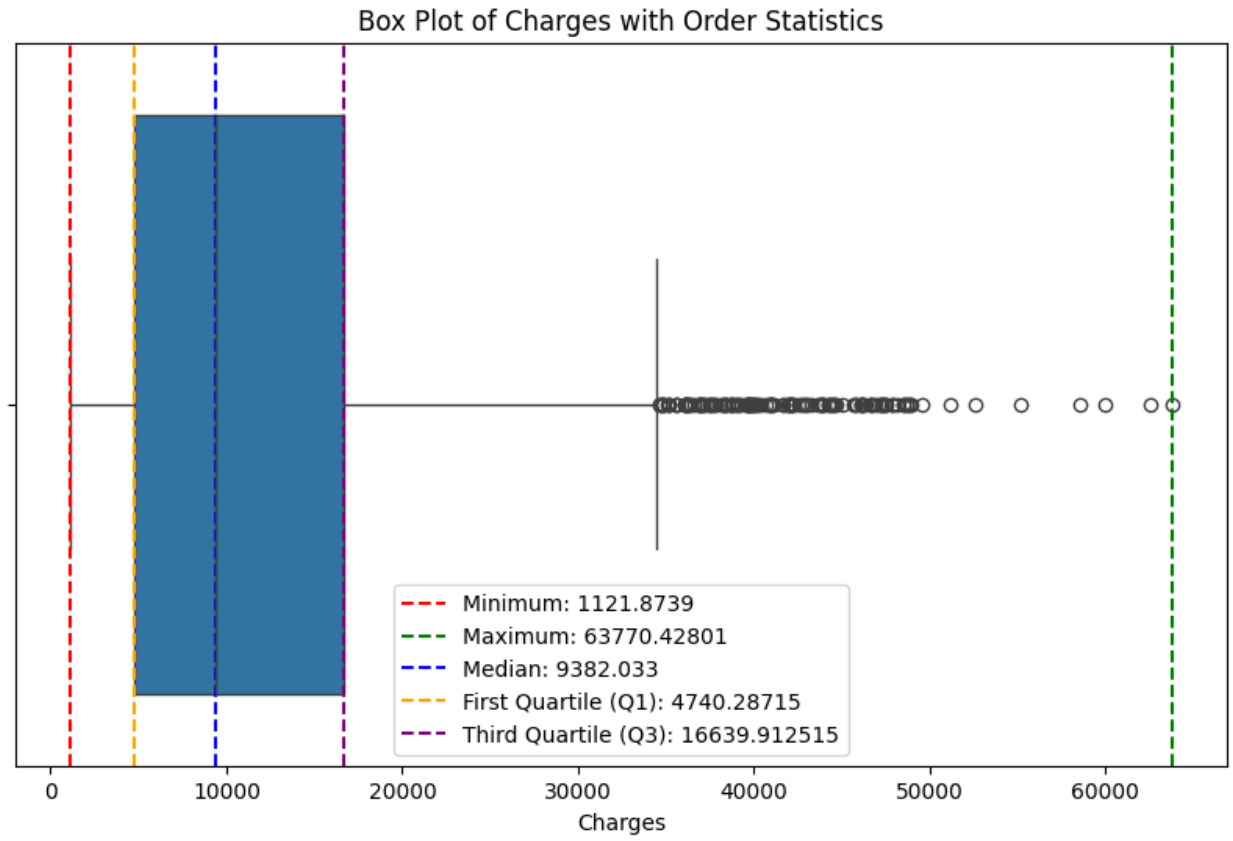


Figure 8: Percentile

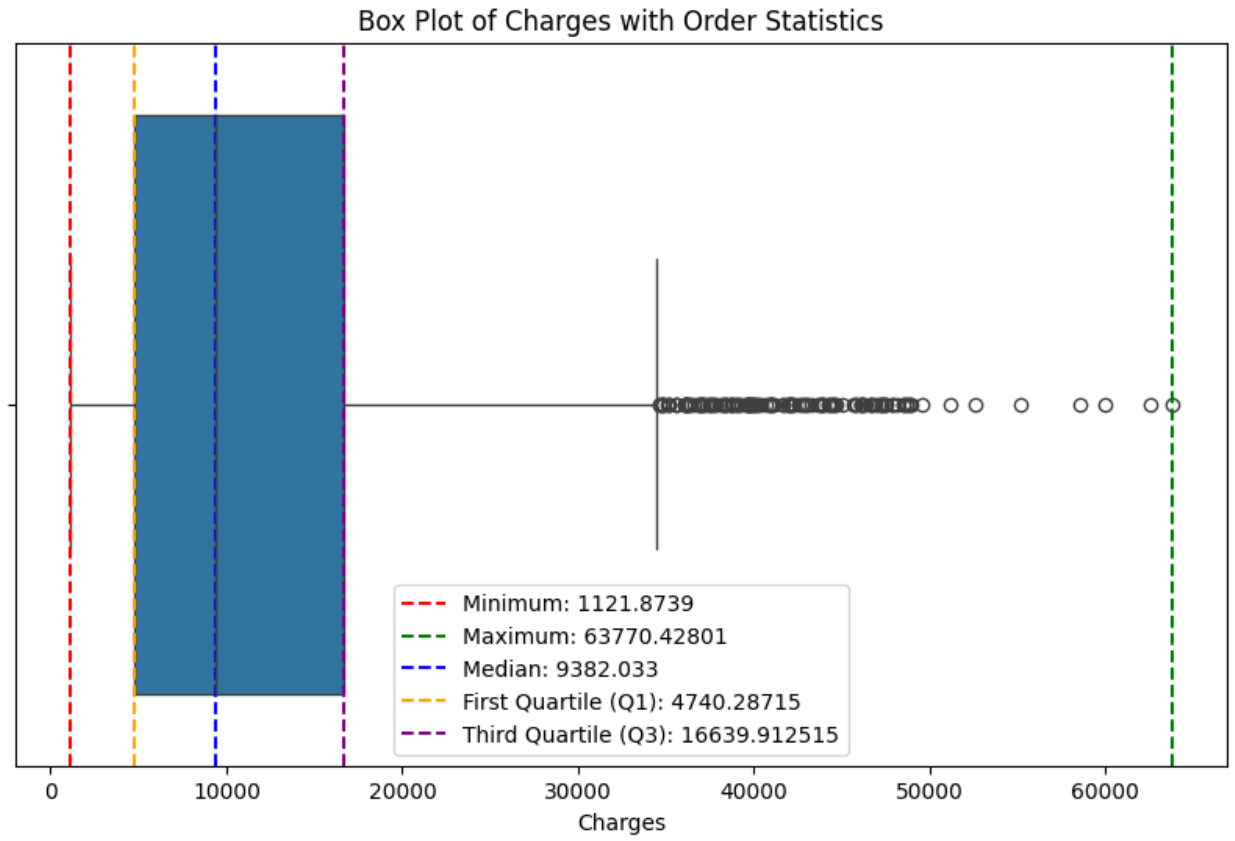


Figure 9: IQR

4.3.1 Key Ideas

- Variance and standard deviation are the most widespread and routinely reported statistics of variability.
- Both are sensitive to outliers.
- More robust metrics include mean absolute deviation, median absolute deviation from the median, and percentiles (quantiles).